AIR QUALITY PREDICTION USING DEEP LEARNING TECHNIQUES

SYNOPSIS

Submitted for the award of the degree of

DOCTOR OF PHILOSOPHY

IN

COMPUTER SCIENCE AND ENGINEERING

by

P. ShreeNandhini(Reg. No. 19PHEOF003)

Ph.D. Research scholar (Full Time)

Under the guidance of

Dr.P.AMUDHA

Professor

Department of Computer Science and Engineering

School of Engineering

Avinashilingam Institute for Home Science and Higher Education for Women

Coimbatore

December 2023

AIR QUALITY PREDICTION USING DEEP LEARNING TECHNIQUES

1. INTRODUCTION

The development of industry and vehicles has resulted in global air pollution, which is a major issue with a bigger impact on human health pollution. In India, air pollution is considered a widespread problem. The situation is worse in large urban areas like Delhi where Air quality Index (AQI) has achieved the highest total time of 999 AQI [11]. An effective system for monitoring and predicting air quality in advance has great importance for human health and government decision-making.Air quality prediction has become a challenging issue due to the complex characteristics with time-space nonlinearity and multi-dimensional feature interaction. The main intention of air quality prediction is to predict the future state of air quality in a particular location based on the existing data, such as meteorological data and historical air quality data.

Various methods have been proposed for prediction of air quality. Machine learning methods such as linear regression neural network are used for air quality prediction. However, the accuracy of these air quality prediction systems is hampered by the complicated array aspects namely conditions of meteorology, emissions and traffic pattern. Also, they suffer from the problems of being prone to local optima and over fitting. Recently, deep learning has been used for air quality prediction which can capture and learn complex relationships among the variables of air quality, leading to more accurate predictions. Deep learning models, particularly neural networks, have shown significant promise in improving the accuracy of air quality predictions. The favorable properties of deep learning make it suitable for modeling and predicting air quality.

The main aim of research work is to implement air quality prediction using deep learning techniques. In this work, deep learning-based algorithms such as Improved Sparse Auto Encoder algorithm, Voronai Clustering Sparse Auto Encoder algorithm, Transferred Stacked Bidirectional and Unidirectional Long Short Term Memory Algorithm, Wasserstein Distance Deep Transfer Learning Algorithm have been proposed to provide a prediction model with good accuracy by performing comparison between the results obtained with these techniques.

2. LITERATURE SURVEY

Tiwari et al. [14] presented a novel deep learning method for long-term air-quality prediction for selected parts of Delhi, India. The deep learning method such as Long Short-Term Memory (LSTMs) networks, Bidirectional-LSTMs and Encoder-Decoder LSTMs have been used. However, the LSTM took longer time to train which provided uncertainty quantification in predictions.

Liang et al. [5] suggested AQI prediction models for acute air pollution in which the algorithm such as random forest, adaptive boosting (AdaBoost), support vector machine, artificial neural network, and stacking ensemble methods were investigated to train models. Moreover, this research observed how prediction performance decayed over longer time frames, and the precision was measured with three commonly used scale-dependent error indexes like mean absolute error (MAE), root mean squared error (RMSE), and R-squared (R2). However, the classification of this approach was ineffective due to high incoming data.

Mao et al. [7] proposed Deep Learning framework using a neural network equipped with a Temporal Sliding LSTM Extended Model (TS-LSTME). Sliding prediction with the right time lag was achieved with the help of hourly historical PM2.5 concentrations, meteorological data, and temporal data, as well as a multi-layer Bi-LSTM. But due to the sophisticated pollutant spread mechanism, there were limitations on how these models can be employed because of problems including the large amount of computational work involved, the complexity of the processes involved, and the uncertainty of the parameters.

Yang et al. [15] proposed a space-time support vector regression model for prediction of hourly $PM_{2.5}$ concentrations. Initially, spatial clustering was implemented to split the study are into various homogenous or quasi-homogenous subareas to address the spatial heterogeneity. Gauss vector weight function was developed to find spatial autocorrelation variables as part of the input features. Finally, a local support vector regression model with spatial autocorrelation variables was established for each subarea.

Zhang et al. (17) suggested hybrid learning strategy in which the original AQI data was decomposed into lower-frequency subseries using Wavelet Packet Decomposition (WPD). The particle swarm optimization approach can be used once the pigeon-inspired optimization has been fine-tuned. The Extreme Learning Machine (ELM) weights and thresholds were optimized using the Improved Pigeon-Inspired Optimization (IPIO) method, and the subseries predictions were made using the Modified ELM (MELM) approach.

3. PROBLEM DEFINITION

Air pollution is a major threat to health and climate. Traditional methods of air quality monitoring rely on sparse networks of stationary sensors, leading to limited spatial coverage and real-time data. Machine learning algorithms can effectively extract and process complex sequences and extract non-linear features. Still, they suffer from the problems of being prone to local optima and overfitting. The prediction model may not learn enough knowledge from the historical data to generate accurate prediction. Therefore, a predictive model which can mitigate the limitation and improve the prediction accuracy at larger resolutions is necessarily needed. In this work, air quality prediction models are built by leveraging the capabilities of deep learning algorithms.

4. OBJECTIVES OF THIS RESEARCH

The main objectives of this research work are to build and train models using deep learning algorithms and find out the most accurate model in predicting the air quality.

- To build predictive deep learning models for air quality prediction.
- To identify the temporal and spatial features and the long-time delay-based locations for better air quality prediction model.
- To minimize the learning from long-term dependencies of air quality data using Transferred Stacked Bidirectional and Unidirectional Long Short-Term memory.
- To optimize the LSTM-based parameters by Wasserstein Distance-Deep Transfer learning.

5. RESEARCH METHODOLOGIES

The dataset chosen for this work is the air quality dataset (https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india) [9] which includes hourly, daily air quality and AQI (air quality index) data from numerous stations in several Indian cities. The data are for the years 2015 through 2020. The original dataset contains 29,532 rows and 16 columns, which included all of the cities: Ahmedabad, Aizawl, Amaravati, Amritsar, Bangalore, Bhopal, Brajrajnagar, Chandigarh, Chennai, Coimbatore, Delhi, Ernakulam, Gurugram, Guwahati, Hyderabad, Jaipur, Jorapokhar, Kochi, Kolkata, Lucknow, Mumbai, Patna, Shillong, Talcher,

Thiruvananthapuram, and Visakhapatnam. The attribute information is: Date YYYY-MM-DD, City, PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, AQI, and AQI_Bucket. AQI_Bucket has six values such as good, satisfactory, moderate, poor, very poor, and severe. The data has been publicly available by the Central Pollution Control Board: https://cpcb.nic.in/ which is the official portal of Government of India.

As air quality is influenced by various factors, including meteorological conditions, emissions from different sources, and geographical features, initially the dataset contained noisy, inconsistent data and missing values. Hence, the data pre-processing was done to handle the unwanted data, duplicated entries and missing values and the data was split into training and testing data. The initial dataset has an imbalanced composition and oversampling technique was used to transform the imbalanced dataset into a balanced dataset. MATLAB R2019a was used as a platform for implementing and experimenting with deep learning algorithms. The system requirement for implementationwas Intel(R) Core™i3-4160T CPU@3.10 GHz 3.09 GHz processor, 4.00 GB RAM, Windows 10.1 pro, 64-bit operating system and 1 TB hard disk.

5.1 Improved Sparse Auto Encoder using Deep Learning Algorithm

In the first part of the research work, an Improved Sparse Auto-encoder with Deep Learning(ISAE-DL) is a proposed approach that uses the feed forward neural network as a sparse autoencoder. The data gathered from diversified nodes with the sensor may comprise noise and anomaly related data. The combined k-Nearest Neighbor Euclidean Distance (kNN-ED) and k-Nearest Neighbor Dynamic Time Warping Distance (kNN-DTWD) is applied to acquire the particulate matter and the meteorological data. Additionally, ANN and LSTM are used to acquire the relative information and the prediction model is generated with training data. To handle the noisy information, an improved sparse auto encoder is introduced, which is feed forward neural network. It effectively handles the noisy information and predicts the air quality accurately. It still cannot handle long delay dependencies so moved to next phase of work.

5.2 Voronoi-Based Clustering Sparse Auto Encoder using Deep learning algorithm

In the second part of the research work, a Voronai-Based Clustering using Sparse Auto Encoder with Deep Learning (VCSAE-DL) has been proposed for addressing the long-time delay-based locations for improved air quality prediction. Then, the temporal and spatial features were identified to retrieve the most important features for air quality prediction. The similar spatially and temporally correlated locations were grouped based on Manhattan distance. The formation of

clusters is continued with different centers and the clustering process is stopped until all the data are covered. The clustered data and the terrain information are given as input to ANN, CNN and LSTM and their results are combined and transferred to sparse auto encoder for the prediction of air quality. It cannot handle Bidirectional and unidirectional LSTM so moved to next phase of work.

5.3 Transferred Stacked Bidirectional and Unidirectional Long Short term Memory Algorithm

In the third phase of research work, an Transferred Stacked Bi-directional and Uni-directional Long Short Term memory (T-SBU-LSTM) has been proposed to minimize the learning issue in long term dependencies for LSTM in the air quality prediction model. Transfer learning is used to improve prediction accuracy at higher temporal resolutions which identifies the similarities between two separate datasets, tasks, or models to transmit data from the source to the new domain. Then, the Transferred Stacked Bidirectional and Unidirectional LSTM (T-SBU-LSTM) has been adopted in learning from long-term PM2.5 dependencies, and it uses Transfer learning to transfer knowledge from smaller temporal resolutions to higher temporal resolutions. This combined architecture enhances the feature learning from the large-scale spatial-temporal time series data by learning both forwards and backward dependencies. In this phase of research, expands the air quality prediction from a specific location to several adjacent locations varying small period to long period time delays.It cannot handle larger size of source and target domain so moved to next phase of work.

5.4 Wasserstein Distance using Deep Transfer Learning Algorithm

In fourth part of research, Wasserstein Distance-Deep Transfer learning (WD-DTL) has been proposed to minimize the learning time of Transfer learning due to larger size data of source and target domains through adversarial training. Initially, a base LSTM model was trained with sufficient data in source domain. Then, Wasserstein distance based DTL (WD-DTL) was constructed to learn invariant features between source and target domains. After this procedure, a discriminator was used to optimize the LSTM-based feature extractor parameters by minimizing the estimated empirical Wasserstein distance. Through the above adversarial learning process, the transferable features from a source domain where faulty labels are known can be brought to diagnose a new but relevant diagnosis task without any labeled sample.

6. EXPERIMENTAL RESULTS AND DISCUSSION

To analyze the performance of the proposed algorithms, the metrics such as Accuracy, Precision, Specificity, Sensitivity, Area Under Curve, Mathew Correlation Co-efficient (MCC) and Mean Absolute Error Rate (MAER)were computed.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Sensitivity = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (3)$$

$$Specificity = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \quad (4)$$

$$AUC = \frac{sensitivity + specificity}{2} \quad (5)$$

$$MCC\ Value = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (6)$$

$$MAER = \frac{FP + FN}{TP + TN + FP + FN} \quad (7)$$

where TP is True Positive, FP is False Positive, FN is False Negative, and TN is True Negative were determined for each algorithm from the confusion matrix. The existing method Spatial Temporal Deep Neural Networks(ST-DNN) was compared with the proposed algorithms of this research work: Improved Sparse Auto Encoder-Deep learning (ISAE-DL),Voronai Clustering Sparse auto encoder algorithm-deep learning(VCSAE-DL),Transferred Stacked Bidirectional Unidirectional Long Short-Term Memory (T-SBU-LSTM) and Wasserstein Distance using Deep transfer learning (WD-DTL)and to interpret the experimental resultant values easily, a graphical representation is given in Figures 1to 6. Figure 1 displays a graphical comparison of all five deep learning-based model in terms of accuracy.

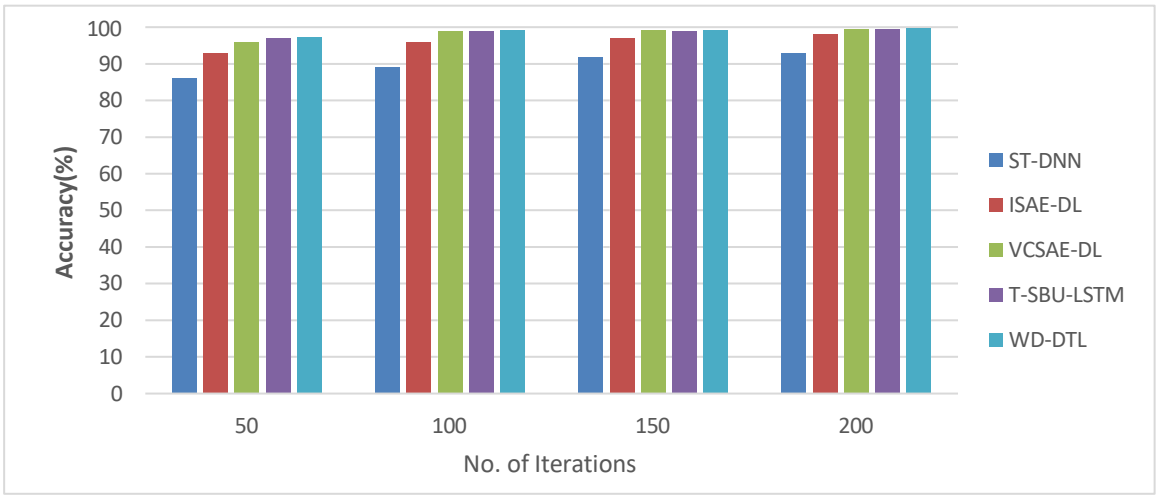

Figure 1. Comparison of Accuracy

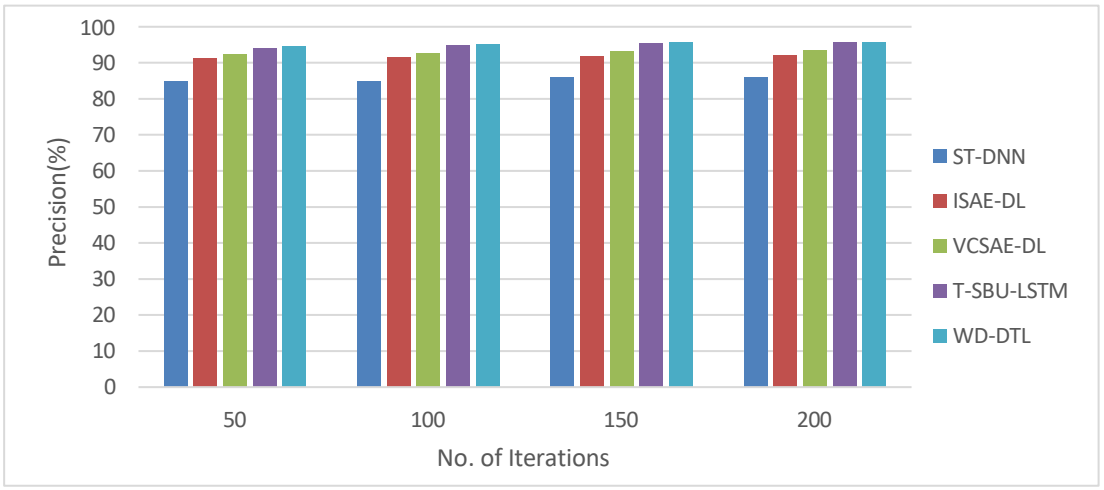

Figure 2. Comparison of Precision

From the Figure 1 it is evident that, the accuracy of ST-DNN is 93%, and the proposed methods ISAE-DL, VCSAE-DL, T-SBU-LSTM and WD-DTL achieves an accuracy of 98%, 99.5%, 99.4% and 99.7% respectively. The performance of ST-DNN, ISAE-DL, VCSAE-DL, T-SBU-LSTM and WD-DTL algorithms are compared based on precision value and is depicted in Figure 2. The computed values of Precision of the algorithms ST-DNN is 86 %, ISAE–DL is 92%, VCSAE-DL is 93.5%, T-SBU-LSTM is 95.6 % and WD-DTL is 95.8%. The outcome of the execution of the proposed models indicates that performance of T-SBU-LSTM is equally comparable with WD-DTL algorithm. WD-DTL model have produced the best performance considering the accuracy and precision values.

As sensitivity and specificity play a crucial role in assessing the overall performance and reliability of air quality prediction models, the sensitivity and specificity values were computed and is given in Figure 3 and Figure 4 respectively.

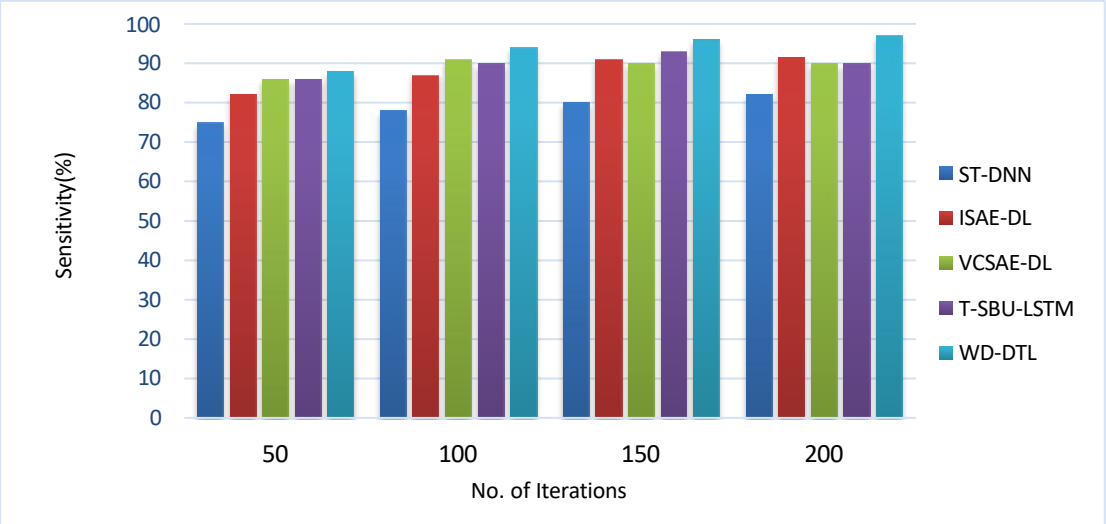

Figure 3. Comparison of Sensitivity

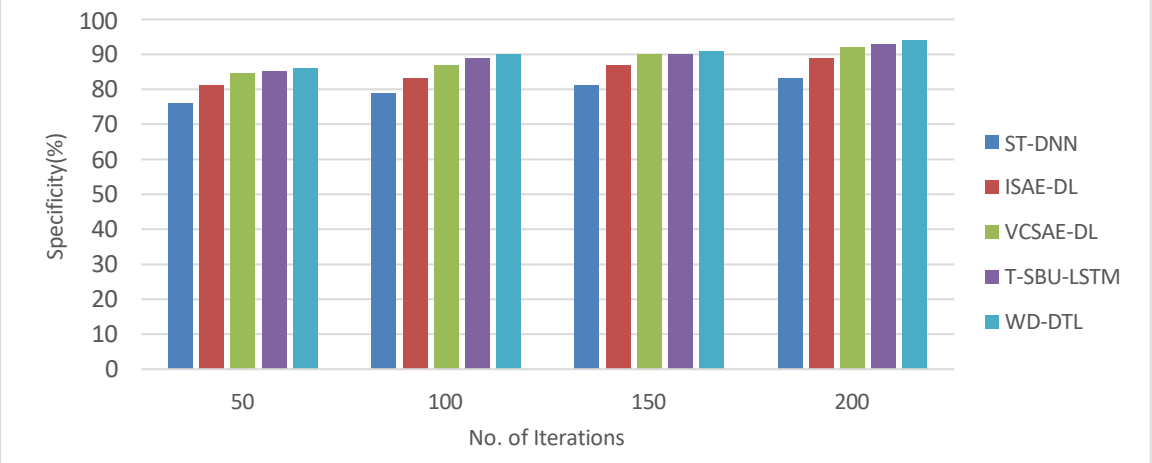

Figure 4. Comparison of Specificity

Figure 3 reveals that the Sensitivity of ST-DNN is 82% whereas, ISAE-DL, VCSAE-DL, T-SBU-LSTM and WD-DTL algorithms obtains 91.5%, 90%, 90%, 97% respectively. From the Figure 4, it is evident that the Specificity of ST-DNN is 83%, and the proposed methods ISAE-DL, VCSAE-DL,T-SBU-LSTM and WD-DTL achieves Specificity values of 89%, 92%, 93% and 94% respectively.Therefore, it is evident from this, that the performance of WD-DTL is comparatively better than other algorithms.

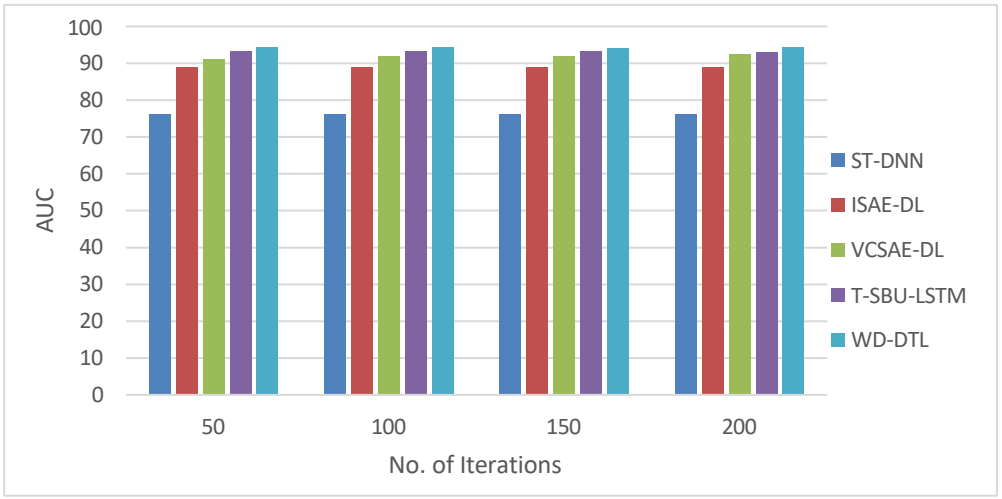

Figure 5. Comparison of AUC

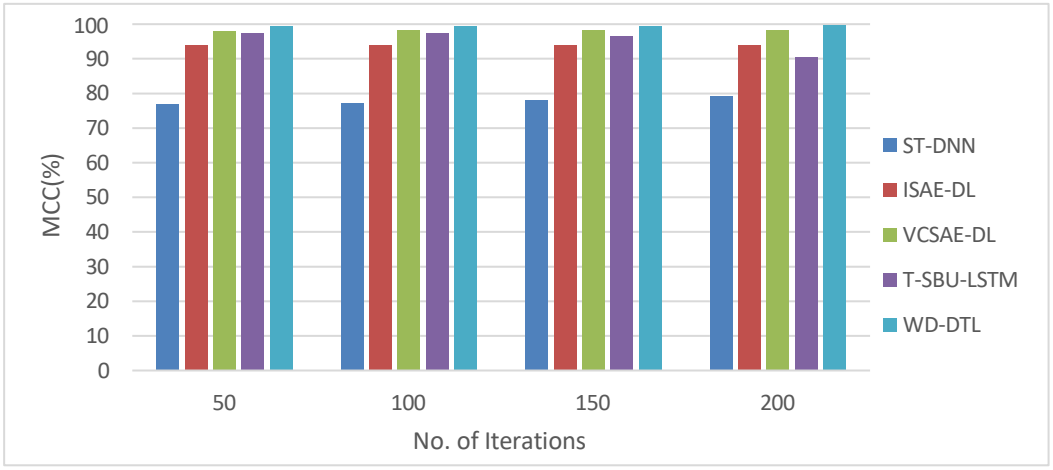

Figure 6. Comparison of MCC

AUC is employed to make the visualization of comparative performance of these five predictive models and is given in Figure 5. It is observed that the AUC score of ST-DNN, ISAE- DL, VCSAE-DL, T-SBU-LSTM and WD-DTL is 0.76,0.88, 0,92, 0.93 and 0.94 respectively. A higher value of AUC refers to more reliability of a predictive model and hence it can be stated that WD-DTL can make reliable predictive model. On comparing the MCC value of all algorithms considered in this work, Figure 6 shows that MCC of WD-DTL model iss calculated to be 99%. This investigation shows that WD-DTL has a higher AUC and MCC than other AQ prediction systems.

The performance of the algorithms is also compared in terms of the statistical measure Mean Absolute Error Rate (MAER) and is given in Figure 7. The y-axis displays the error rates, while the x-axis displays the total number of training epochs. It is observed that the result of the WD-DTL

is better in terms of the evaluation parameter where, MAER =1.4 compared to the results of other algorithms.

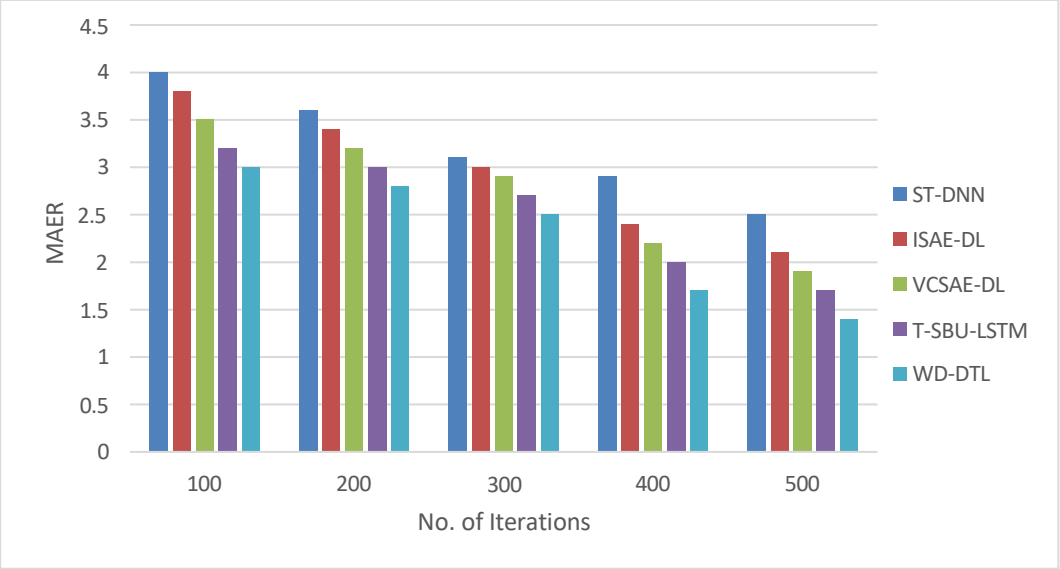

Figure 7. Comparison of MAER

7. CONCLUSION

Air pollution is a global problem where the researcher from all around the world is working to discover a solution. Several machine learning techniques have been proposed for air quality. but do not yield a reliable result for making long-term forecasts of air quality. Deep learning models have the ability to automatically learn intricate patterns and representations from complex data makes them well-suited for tasks where traditional models may struggle. Hence, deep learning based predictive models have been developed for air quality prediction using the benchmark air quality dataset of various Indian cities obtained from Kaggle. The noisy and anomaly related data have been handled and applied the relative information using Improved Sparse Auto Encoder Algorithm. The similar spatially and temporal correlated locations were grouped using Manhattan distance. The long time delay based locations have been addressed using VCSAE-DL. In order to minimize the learning issue in long term dependencies for LSTM, transfer learning method, (T-SBU-LSTM) was used which is the transferred to sparse auto encoder for air quality prediction. Also the learning time of transfer learning was minimized using WD-DTL algorithm. Experiments show that the proposed WD-DTL method outperformed other existing methods in terms of accuracy, precision, specificity, sensitivity, AUC, MCC, and MAER.

8 ORGANIZATION OF THE THESIS

Chapter 1 discusses about the deep learning techniques, Air quality and Contribution of this research work. **Chapter 2** studies and analyses the literature surveys which are related to this research work. **Chapter 3** describes the first contribution of research work such as Improved Sparse auto Encoder Deep learning Algorithm. **Chapter 4** describes the second contribution of the research work such as Voronoi-Based Clustering Sparse Auto Encoder Deep learning algorithm. **Chapter 5** describes the third contribution of the research such as Transferred stacked Bidirectional and Unidirectional Long short term Memory Algorithm **Chapter 6** describes the fourth contribution of the research such as Wasserstein Distance using Deep transfer learning Algorithm. **Chapter 7** explains the complete performance analysis of the proposed methods. **Chapter 8** concludes the research work with a summary of the findings and provides the possibility of future work for further development.

REFERENCES

1. A. Deepak, Dr. Amrapali S. Chavan, Aniruddha Bodhankar, Advancing Air Quality Prediction in Specific Cities Using Machine Learning, International Journal of Intelligent Systems and Applications in Engineering, Volume 11, Issue 11, 2023.
2. Abraham, S., & Li, X., A Cost-effective Wireless Sensor Network System for Indoor Air Quality Monitoring Applications, Elsevier, Volume 34, Issue 1, 2014.
3. Eslami, E., Choi, Y., Lops, Y., & Sayeed, A Real-Time Hourly Ozone Prediction System Using Deep Convolutional Neural Network, Neural Computing and Applications, Volume 32 Issue 1, 2019.
4. Ishan Ayus, Narayanan Natarajan, and Deepak Gupta, Comparison of Machine Learning and Deep Learning Techniques for the Prediction of Air Pollution: A Case Study from China, Asian Journal of Atmospheric Environment, Volume 17, Issue 4, 2023.
5. Li, M., Wang, W. L., Wang, Z. Y., & Xue, Y., Prediction of PM2.5 Concentrations Based on the Similarity in Air Quality Monitoring Network, Building and Environment, Volume 137, Issue 1, 2018.
6. Liang, Y. C., Maimury, Y., Chen, A. H. L., & Juarez, J. R. C., Machine Learning-Based Prediction of Air Quality, Journal on Applied Sciences, Volume 10, Issue 24 ,2020.
7. Ma, J., Cheng, J. C., Lin, C., Tan, Y., & Zhang, Improving Air Quality Prediction Accuracy at Larger Temporal Resolutions Using Deep Learning and Transfer Learning Techniques, Journal on Atmospheric Environment, Volume 214, Issue 1, 2019.
8. Manuel Méndez, Mercedes G, Merayo, Machine Learning Algorithms to Forecast Air Quality: A Survey Journal of Artificial intelligence Review, Volume 56, Issue 1,2023.
9. N. Srinivasa Gupta, Yashvi Mohta, Prediction of Air Quality Index Using Machine Learning Techniques, A Comparative Analysis, Journal of Environment Health, Volume 23, Issue 1, 2023.
10. Prathapa Dishanth, Air Quality Predictive Analysis Using Machine Learning, International Research Journal of Modernization in Engineering Technology and Science, Volume 4, Issue 5, 2022.
11. Rajat R. Relkar, Vaibhav Deulkar , Ridam Gunjewar, Rahul Panghate, Pratik Gaurkar , Mukul Singanjude, RitikJarile , Prof. Preetee K. Karmore, Prediction of Air Quality Index Using Supervised Machine Learning, International Journal for Research in Applied Science & Engineering Technology, Volume 10, Issue VI, June 2022.
12. Soh, P. W., Chang, J. W., & Huang, J., Adaptive Deep Learning-Based Air Quality Prediction Model using the most relevant Spatial-Temporal Relations, IEEE Access, Volume 6, Issue 1, 2018.

13. Syafei, A. D., Fujiwara, A., & Zhang, J. Prediction Model of Air Pollutant Levels Using Linear Model with Component Analysis, International Journal of Environmental Science and Development, Volume 6, Issue 7, 2015.
14. Tiwari, A., Gupta, R., & Chandra, R. Delhi Air Quality Prediction using LSTM Deep Learning Models with a focus on COVID-19 lockdown, Journal of Environment Modeling, Volume 23, Issue 1, 2021.
15. Yang, W., Deng, M., Xu, F., & Wang, H., Prediction of Hourly PM2. 5 Using a Space- Time Support Vector Regression Model, Journal of Atmospheric Environment, Volume 181, Issue 1, 2018.
16. Zhang, L., Liu, P., Zhao, L., Wang, G., Zhang, W., & Liu, J, Air Quality Predictions with a Semi-Supervised Bidirectional LSTM Neural Network, Journal on Atmospheric Pollution Research, Volume 12, Issue 1, 2020.
17. Zhang, Y., Wang, Y., Gao, M., Ma, Q., Zhao, J., Zhang, R., & Huang, L, A Predictive Data Feature Exploration-Based Air Quality Prediction Approach. IEEE Access, Volume 7, Issue 1, 2019.

LIST OF PUBLICATIONS

1. P.ShreeNandhini,Dr.P.Amudha, “Air Quality Prediction Using Voronai-Based Spatial Temporal Sequence Similarity With Conjugate Gradient Enabled Sparse Auto Encoder Deep Learning”,International Journal on Recent and Innovation Trends in Computing and Communication, Vol 11 ,Issue 10,pp 1435-1443 November 2023.(Scopus Indexed)
2. P.ShreeNandhini,Dr.P.Amudha, ”Exploitation of Advanced Deep Learning Methods and Feature Modeling for Air Quality Prediction”, Revue intelligence Artificielle, Vol 36, Issue No 6, pp-959- 967,December2022(Scopus Indexed)

LIST OF CONFERENCES

1. A Review in Air Quality Prediction Using Artificial Intelligence Techniques in Machine Learning Approaches, International Conference Emerging Trends in Industry 4.0 and Sustainable Concepts on 24.2.2023 and 25.02.2023 Organized by Avinashilingam institute for Home Science and Higher Education for Women, Coimbatore.

2. Comparative Analysis of Air Quality Prediction Using Artificial Intelligence Technique, International Conference on Technologies for Smart Green Connected Societies on 29.11.2021 to 30.11.2021 Organized by Avinashilingam institute for Home Science and Higher Education for women, Coimbatore.

3. Air quality Prediction using Artificial Intelligence-A Survey, International Conference on Modern Engineering, Science and Technology on 17.02.2020 Organized by Dr.Kalam Institute of Engineering, Chennai.